# Ethics and Controversies in Natural Language Processing

Jennifer Williams

Edinburgh PhD Student
Centre for Speech Technology Research (CSTR)
j.williams@ed.ac.uk

November 8th, 2018 @ ANLP Lecture

# A LITTLE ABOUT ME

| | |
|---|---|
| **2006-2009** | BA in Applied Linguistics (magna cum laude) from Portland State University |
| **2007** | Museum educator in physics and laser lab |
| **2009-2012** | MS in Computational Linguistics from Georgetown University |
| **2010** | Language Engineer at translation company, LingoSystems |
| **2011-2012** | Developer in R&D at startup company, OpenAmplify |
| **2012** | A*STAR visiting scholar at I2R in Singapore |
| **2012-2017** | MIT technical staff |
| **2017-present** | Edinburgh Informatics PhD student (adversarial learning for speech synthesis) |

Past and current research areas/interests

**Cross-language search**
**Automatic summarization**
**Chinese/English translation**
**Language identification**
**Sign language processing**

**Sentiment analysis**
**Second-language learning**
**Speech pronunciation feedback**
**Text-to-speech Synthesis**
**Signal tampering detection**

# OUTLINE

- What is ethics for NLP
- Why is ethics important
- Who are the stakeholders
- Examples from speech research
- Examples from the workplace

# OUTLINE

- What is ethics for NLP
- Why is ethics important
- Who are the stakeholders
- Examples from speech research
- Examples from the workplace

# INTRO: WHAT DOES ETHICS MEAN?

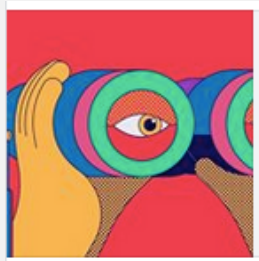
- Merriam-Webster dictionary:
  - The discipline dealing with what is good and bad and with moral duty and obligation
- Association for Computing Machinery (ACM)
  - Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good.
- Also includes:
  - What you do "when no one is looking"
  - Doing the right thing (even if it's hard)
  - Following established laws

Actively seeking to understand the full spectrum of potential consequences, good or bad, of working with algorithms, data, and its impact on self, societies, people, and institutions

# ETHICS IS…

- What you **DO**
- Not only what you believe, think, prefer, or like
- Active behaviors
- Developing as technology develops

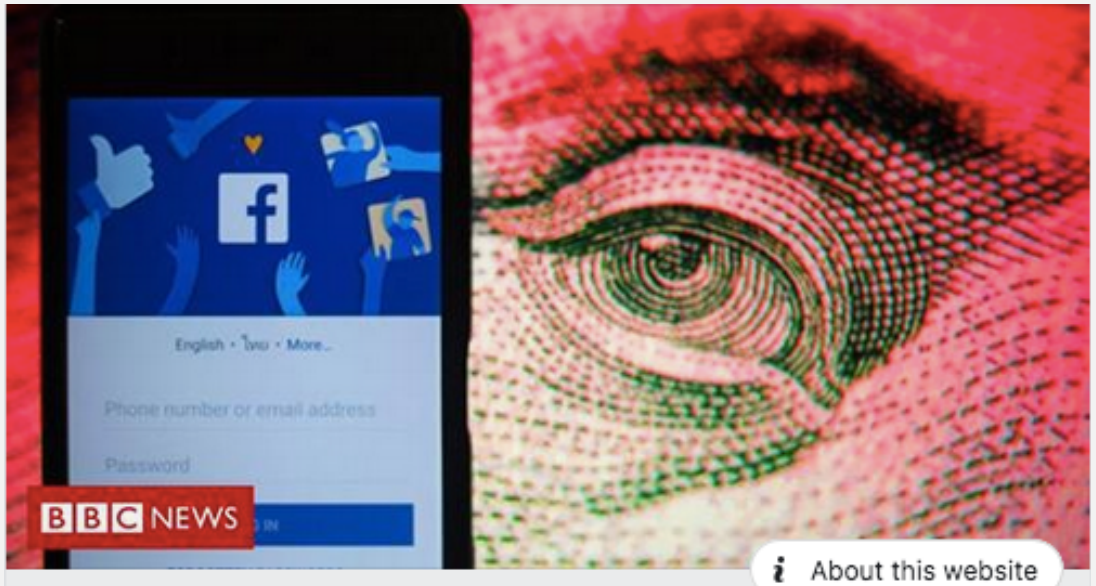
# INTRO: WHAT IS ETHICS IN NLP?



YOUTUBE.COM

**The Truth About Algorithms | Cathy O'Neil**

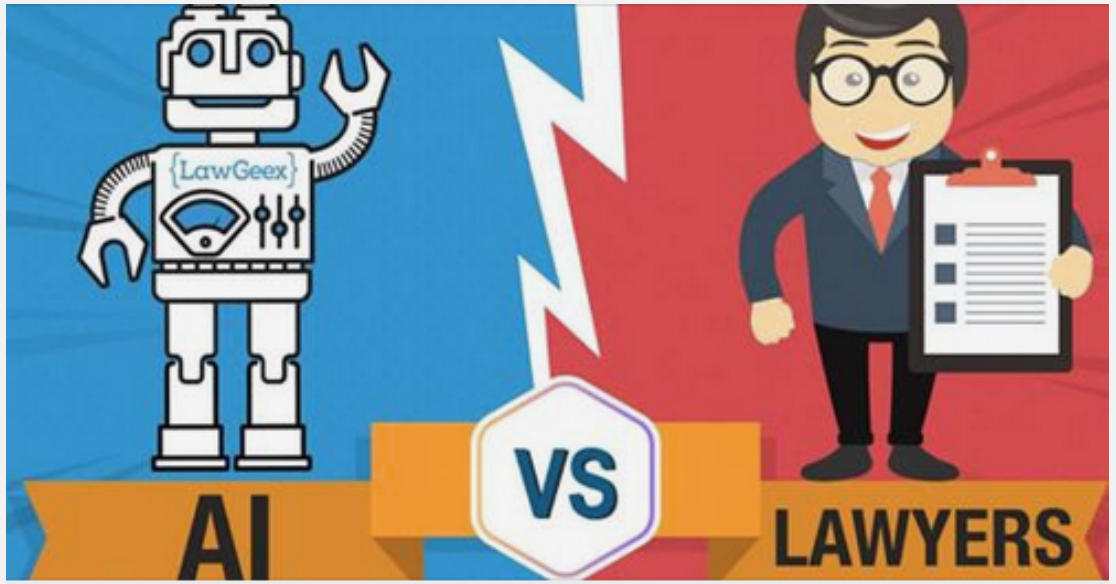
We live in the age of the algorithm - mathematical models are sorting our job applications, curating our online worlds, influencing our elections, and...




BBC.COM

**Hacked Facebook private messages for sale**

The perpetrators told the BBC Russian Service they had details from a...




TECHSPOT.COM

**Machine-learning algorithm beats 20 lawyers in NDA legal analysis**

# INTRO: WHAT IS ETHICS IN NLP?


BBC.COM

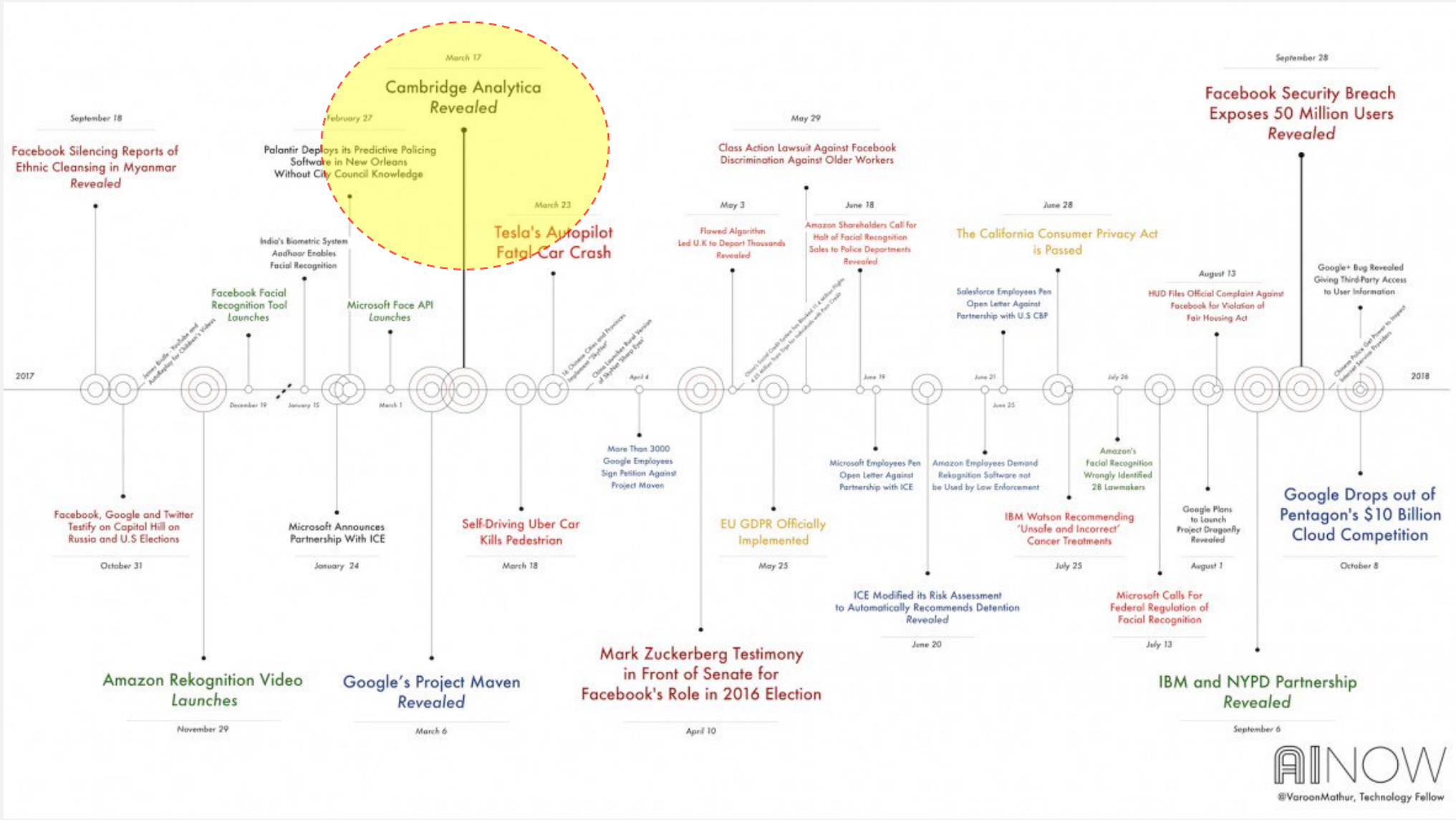
**Under-5s apps have 'unfair deceptive ads'**

Campaign groups and university researchers raise concerns about the...

PSYPOST.ORG

**Study finds link between psychopathy and using Tinder while in a committed relationship**

# ONE YEAR OF ETHICS AND SCANDALS IN AI / NLP 2017 TO 2018



# OUTLINE

- What is ethics for NLP
- Why is ethics important
- Who are the stakeholders
- Examples from speech research
- Examples from the workplace

# WHY DO WE STUDY ETHICS IN NLP?

- Our goal is to give you a compass, and then it's up to you to navigate and made ethical decisions
- More than telling the right technical story
- More than being truthful on assignments, publications, and in the workplace
- Our work in AI / NLP has real consequences
- Ethics includes data, algorithms, analysis, findings, and more
- Laws are different in different countries, and they change and adapt to fit technological landscape

# ETHICAL ISSUES FOR PEOPLE IN NLP

- **Technology-Level Issues**
  - Explainable AI (XAI)
  - Model bias
- **Data-Level Issues**
  - GDPR (May 2018)
  - Internal Review Board & Human Subjects
  - Data fusion, anonymization, and subject withdrawals
- **Career-Level Issues**
  - Published papers retracted/redacted
  - Working on cross-disciplinary teams (lawyers, C-suite, non-scientists, HR, etc)
  - Conscientious objection

# OUTLINE

- What is ethics for NLP
- Why is ethics important
- Who are the stakeholders
- Examples from speech research
- Examples from the workplace

# WHO ARE THE STAKEHOLDERS?

- Science itself
- Companies & Institutions: boss, CEO, shareholders, clients
- Society: laws, individuals, [vulnerable] groups, quality of life
- You: degree, job/career, family, legacy, reputation
- Governments/nations: different laws, cultures, customs, beliefs
- Anyone you will have to explain your work to (non-technical audience)
- what is conscientious objection?

# WHO THINKS THIS INVOLVES ETHICS…

**Collect some Twitter data from the API, determine if a twitter user is suicidal or not**

**Ask humans to label sentiment for news articles about mass murder**

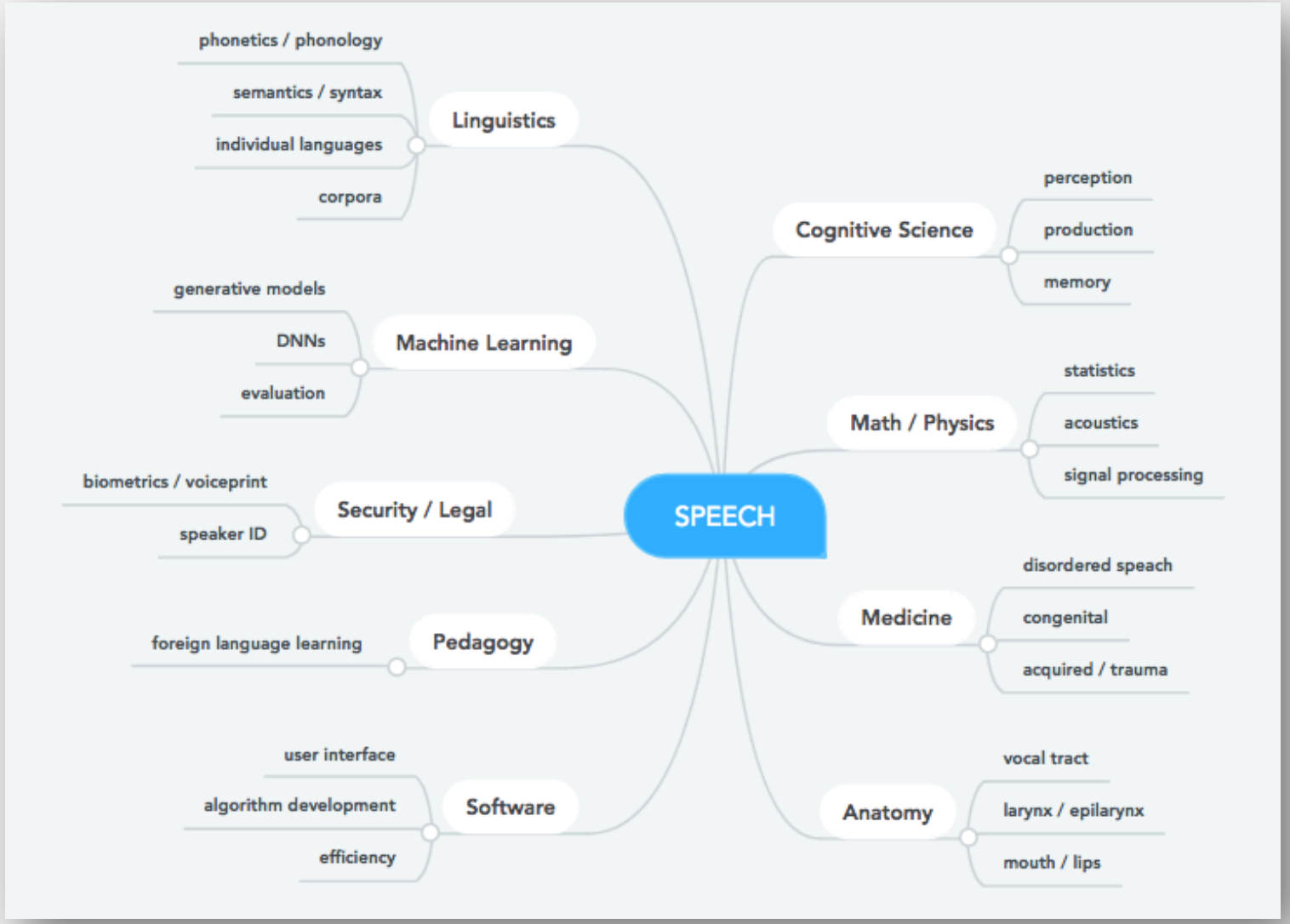
**Build a dialect model using Wall Street Journal data**

**Translate news articles into a regional dialect (Brazilian Portuguese)**

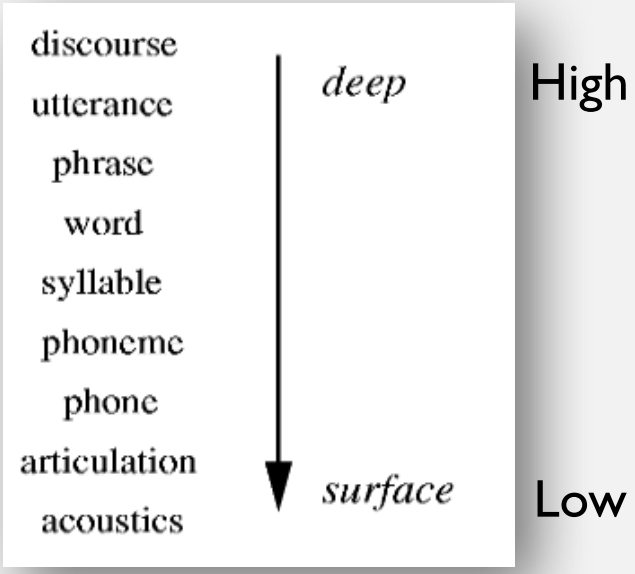
# OUTLINE

- What is ethics for NLP
- Why is ethics important
- Who are the stakeholders
- Examples from speech research
- Examples from the workplace

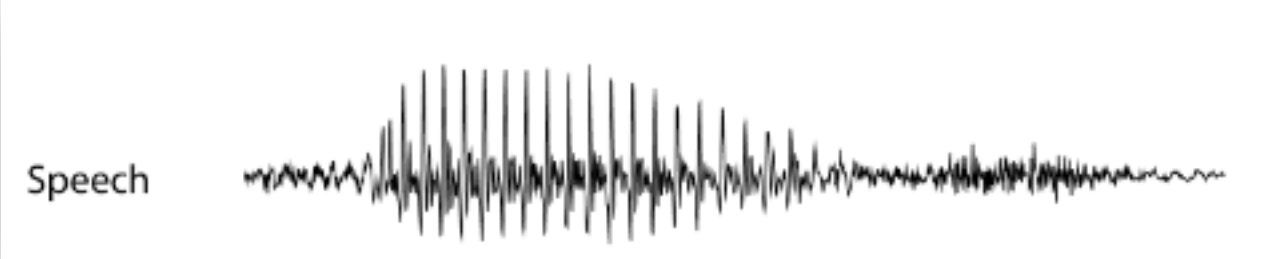
# SPEECH AND NLP IS INTERDISCIPLINARY



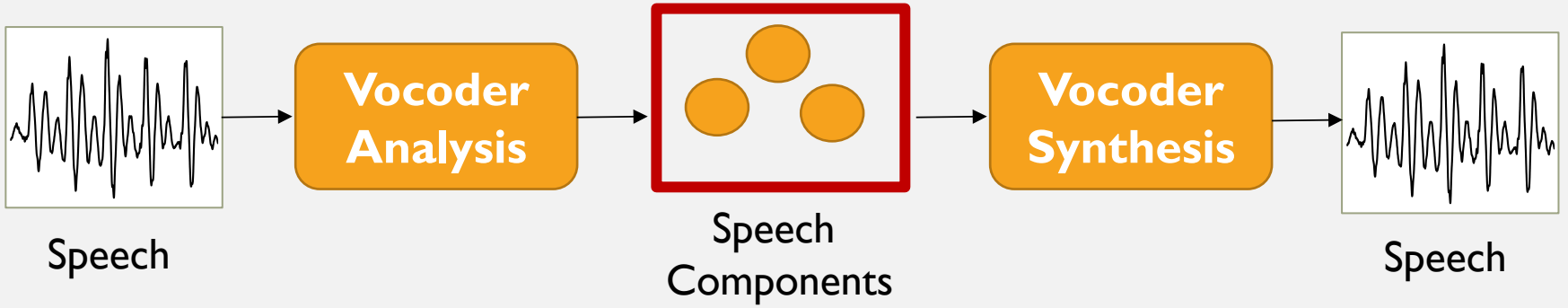

Levels of Abstraction



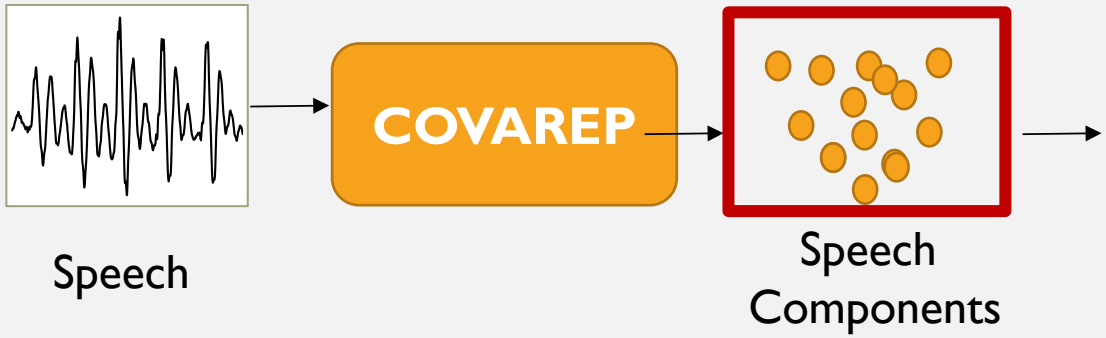
# SPEECH PROCESSING




Pull apart the signal in lots of different ways
Signal components vary by task
Many mathematical transforms on the signal
Put the signal back together




"Copy Synthesis"




Create feature vectors for machine learning
Modify individual components of the speech signal
Study the role of individual signal features
Identify redundant information (compression)

# VOICEPRINT TECHNOLOGY

- Voiceprint = measurable characteristics of the speech signal that identify an individual
- Developing since ~ 1985/90

Size, shape of vocal tract filter

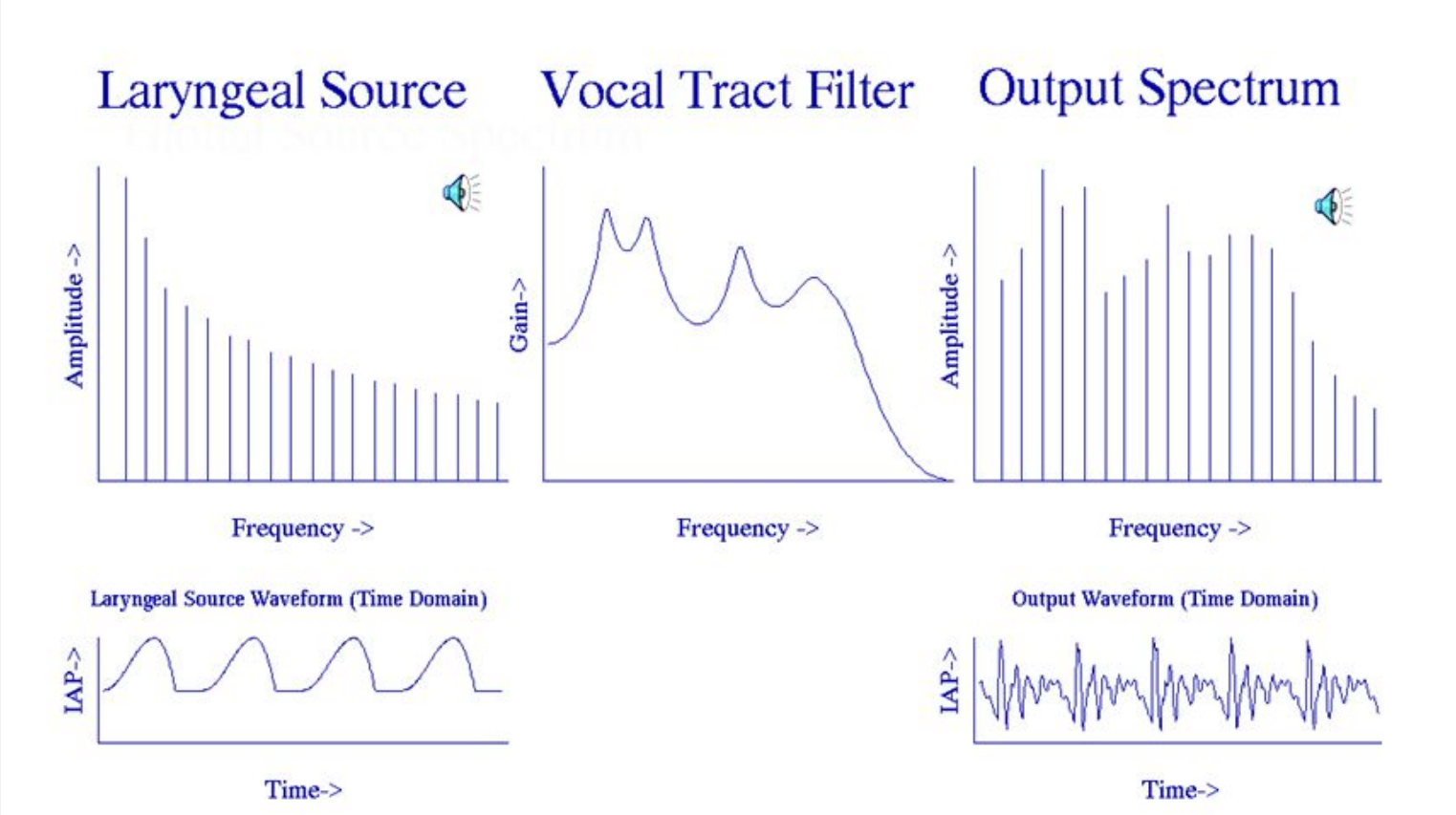
Height/weight

Gender

Native language (if accented)

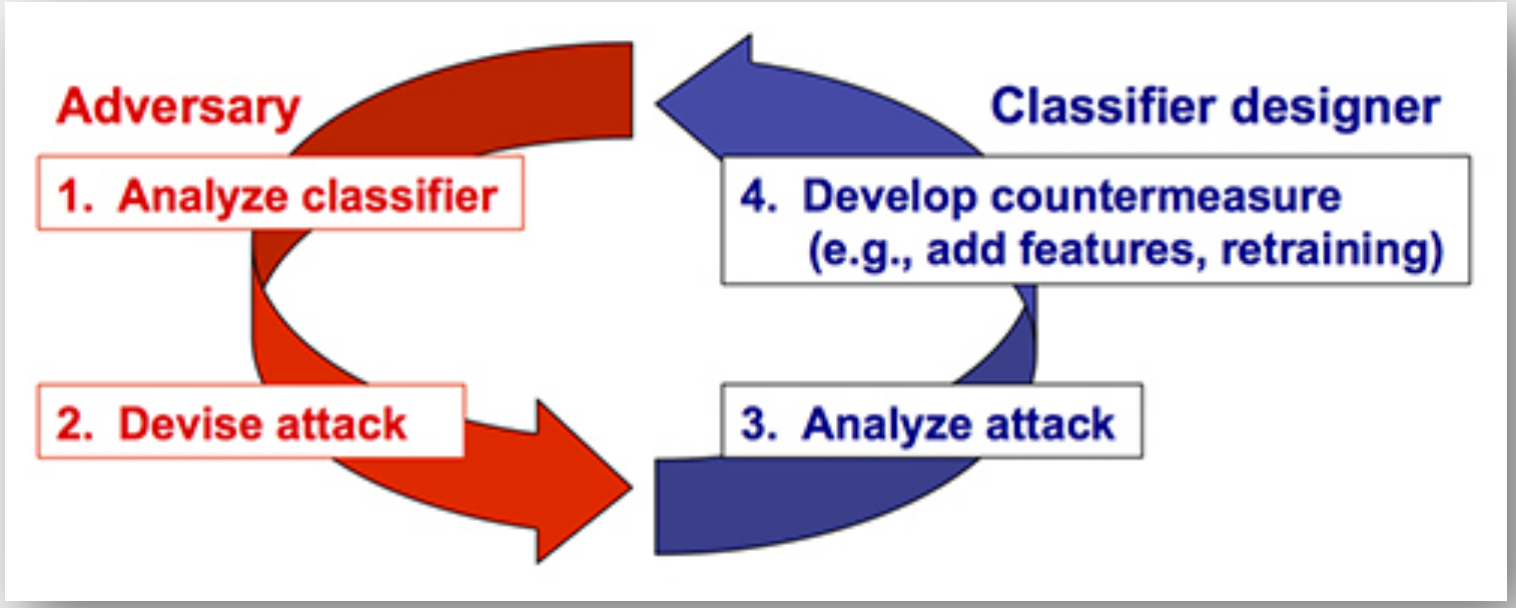
Native origin region

Age



# SPOOFING ARMS RACE

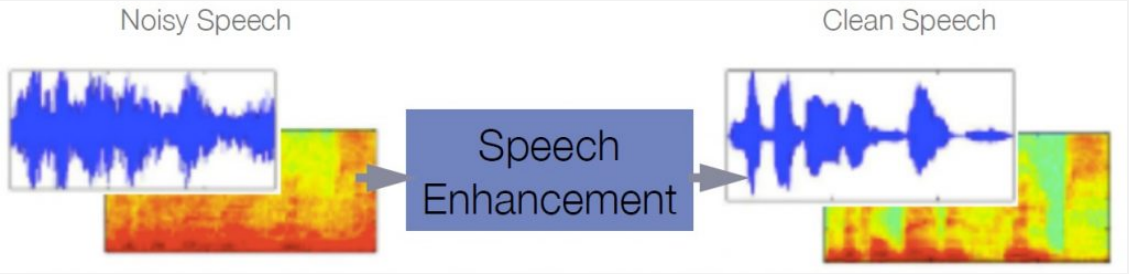
1) Speaker ID
2) Impersonate (spoofing)
3) Speaker ID + anti-spoofing
4) Better spoofing
5) Better anti-spoofing
6) Spoofing + anti-spoofing
   technology developed
   side-by-side
7) International Spoofing Challenges
8) Tech transfer to other voice technology



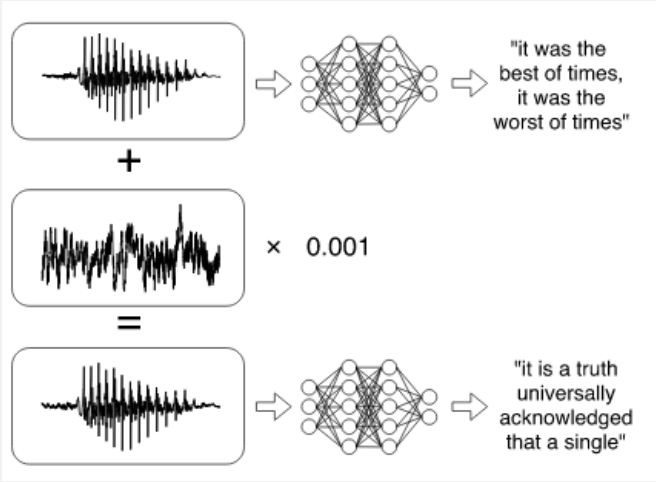

**A normal and academically acceptable approach to speech technology R&D**

# SPEECH SIGNAL MODIFICATION

- Enhancement
- Watermarking / steganography
- Noise reduction




- Covert signal embedding (dolphin attack)
- Speech-enabled device hi-jacking
- Voiceprint spoofing




**Similar technology is used for attack and non-attack**

# WHO THINKS....

**It is wrong to study adversarial techniques in speech and NLP**

**It is OK to study adversarial techniques in speech and NLP**

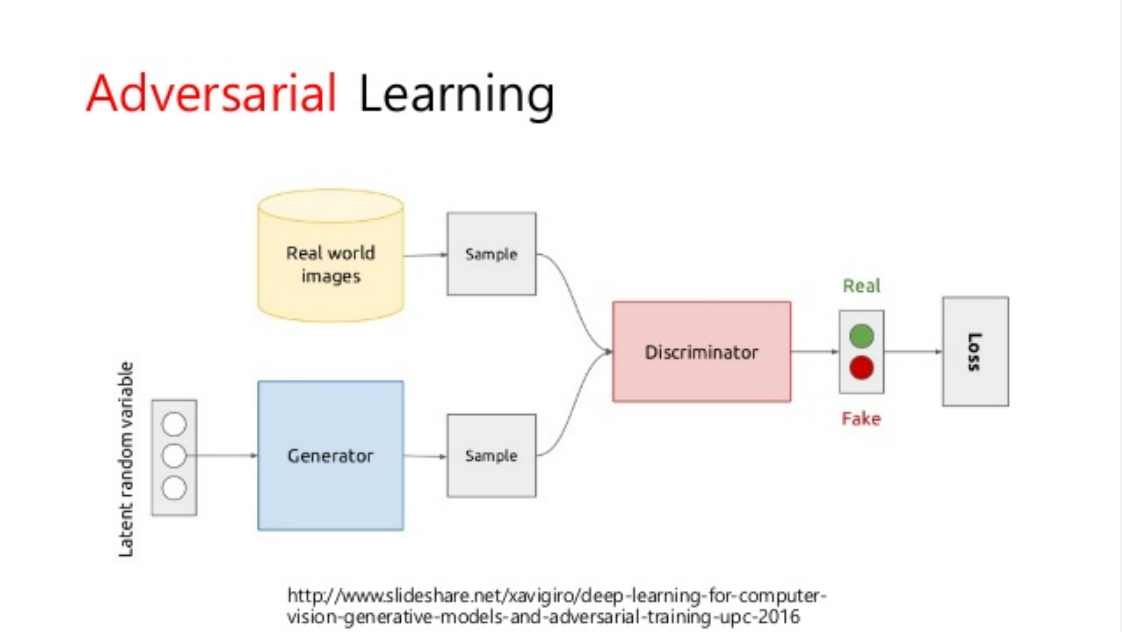
**It depends, I'm unsure, I need to think about this more**

# ATTACK AND HACK

- Fishing attacks (can you hear me, "yes", then splice the audio)
- Impersonation attack (machine to machine) – speaker ID
- Impersonation attack (machine to human) – Google personal assistant
- Replay attack: pre-recorded speech
- Dolphin attack: ASR inaudible voice commands
  - https://www.youtube.com/watch?v=w0Gq5JqC_ts
- Dolphin attack: ASR concealed voice commands
  - https://nicholas.carlini.com/code/audio_adversarial_examples/
- Internet of Things attacks

# SPEECH SYNTHESIS RESEARCH (FOR MY PHD)

- Continue working on TTS synthesis for PhD
- GANs
  - Condition for TTS
  - Speaking style
  - Speaker identification
  - Prosody
- Bridge the gap between learning/modeling and user application
  - Machine learning results can show limits and possibilities
  - How does this reach the user?
  - Allow user to modify their speech just as a human does (audience, comprehension, emotion, etc)



# OUTLINE

- What is ethics for NLP
- Why is ethics important
- Who are the stakeholders
- Examples from speech research
- Examples from the workplace

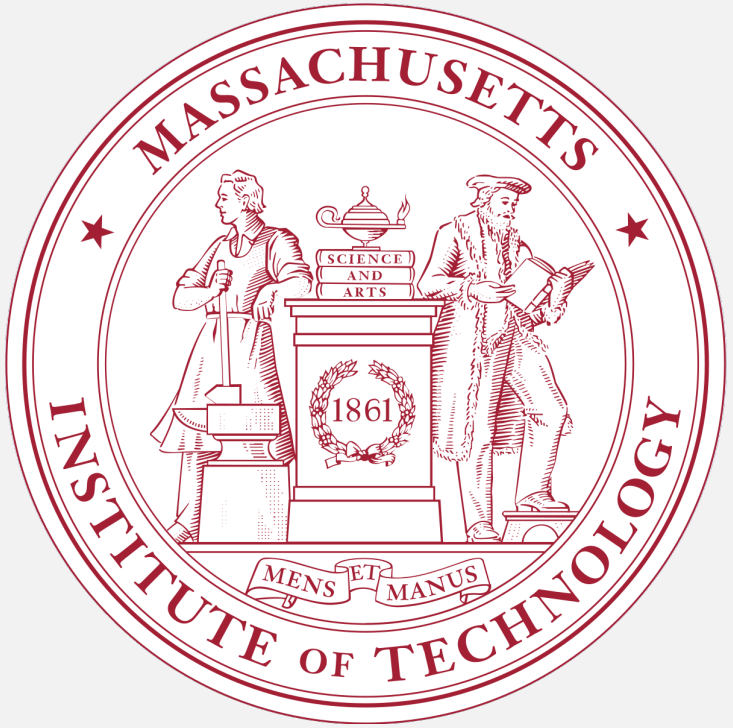
# EXAMPLES FROM THE WORKPLACE

- Employee expectations
- Employer expectations
- Unknown unknowns (you may need a compass to navigate this)

# EXAMPLE #1 – FOLLOWING COPYRIGHT LAWS

- MIT, Cambridge Massachusetts (USA)

- Crawl news RSS feeds in various languages
- Determine reading level of each story
- Relay story + reading level to clients
- Copyright issues raised by client
- In-house legal department
- 4-5 weeks collaboration
- Describe what the algorithm does
- Final determination: no copyright infringement
- 10-page dossier generated for any future inquiries



# EXAMPLE #2 – ANALYZING OBJECTIONABLE MATERIAL

- Alan Turing Institute (London)
- UK Cabinet Office (Defence and Security)
- Sign an NDA
- Agree to take breaks from looking at materials
- Analyze ISIS propaganda
  - Terrorist instructions, religions doctrine, ideology
- Ad-hoc team of varying backgrounds and abilities
- Use NLP to gain insights
- Write a report
- NDA: do not discuss methods, techniques, findings, or content of propaganda

The Alan Turing Institute

Our mission as the national institute for data science and artificial intelligence is to make great leaps in research in order to change the world for the better.

# EXAMPLE #3 – RECRUITING HUMAN JUDGEMENTS

- A*STAR Institute for Infocomm Research, Singapore
- Chinese-English simultaneous translation
- IRB approval at National University of Singapore
- Universal Declaration of Human Rights
- English, Chinese, Spanish
- Identify "units of meaning" in each text
- Outline experiment beforehand
- IRB approval included analysis of potential adverse effects on humans
- Important for publication

SINGAPORE Agency for Science, Technology and Research

# THANK YOU

## Additional Resources

Cathy O'Neil, 2016. Weapons of Math Destruction PDF free online

Cathy O'Neil short YouTube video on algorithms and bias: https://bit.ly/2QkFYz6

Ethics in NLP Wiki page: https://aclweb.org/aclwiki/Ethics_in_NLP


Goodman, B., & Flaxman, S. (2016). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *arXiv preprint arXiv:1606.08813*.

Nissenbaum, H. (2004). Privacy as Contextual Integrity. Washington Law Review 79(1): 119–58.

Ohm, P. (2009). Broken promises of privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, *57*, 1701.

Solan, L. M., & Tiersma, P. M. (2002). Hearing Voices: Speaker Identification in Court. *Hastings Law Journal*, *54*, 373.

Yakowitz, J. (2011). Tragedy of the Data Commons. *Harvard Journal of Law & Technology*, *25*, 1.